Classification-oriented structure learning in Bayesian networks for multimodal event detection in videos

Guillaume Gravier, Claire-Hélène Demarty, Siwar Baghdadi, Patrick Gros

Classification-oriented structure learning in Bayesian networks for multimodal event detection in videos

Guillaume Gravier *

CNRS – IRISA, Rennes, France

Claire-Hélène Demarty

Technicolor, Rennes, France

Siwar Baghdadi

Technicolor, Rennes, France

Patrick Gros

INRIA, Rennes, France

* *Corresponding author address:* Guillaume Gravier, CNRS – IRISA, Campus de Beaulieu, 35042 Rennes Cedex, France.

E-mail: guillaume.gravier@irisa.fr

ABSTRACT

We investigate the use of structure learning in Bayesian networks for a complex multimodal task of action detection in soccer videos. We illustrate that classical score-oriented structure learning algorithms, such as the K2 one whose usefulness has been demonstrated on simple tasks, fail in providing a good network structure for classification tasks where many correlated observed variables are necessary to make a decision. We then compare several structure learning objective functions, which aim at finding out the structure that yields the best classification results, extending existing solutions in the literature. Experimental results on a comprehensive data set of 7 videos show that a discriminative objective function based on conditional likelihood yields the best results, while augmented approaches offer a good compromise between learning speed and classification accuracy.

1. Introduction

Automatic video analysis is one of the basic tools that enables the development of applications such as multimedia information retrieval or novel TV services. Understanding a video is a multifaceted objective that includes simple tasks like the recovery of the video structure or more complex ones such as the detection of specific events. In this paper, we focus on this last case, i.e., on the ability to detect specific extracts of a video that have a particular, and usually important, meaning for the user. Event detection is particularly important when videos have a weak structure since events are then the only anchors in the stream that allow non-linear browsing.

Events have no general definition and are in general specific to a particular context and application. Most of the time, the application context calls for a high-level semantic definition of an event (a goal in soccer, a dunk in basketball), but the definition may be much fuzzier: An interesting moment, or a moment similar to a set of examples. This suggests two main approaches to build an event detection system. First, a formal definition of the event can be given as a model which is then used for detection. Such approaches are usually based on rules that can be either defined by a human expert (see, e.g., Saraceno and Leonardi (1998), Tovinkere and Qian (2001), Lienhart et al. (1998), Zhong and Chang (2001)) or inferred from examples as in Perlovsky (1998). Alternately, machine learning techniques can be used to train a system from examples with the goal of deciding whether a video extract contains the event or not. These approaches heavily rely on classification techniques, be they probabilistic (see, e.g., Huang et al. (1999), Xu et al. (2002), Mizutani et al. (2005)) or not (e.g., Bae et al. (2005), Snoek and Worring (2005), Haering et al. (2000)). The goal

of these classifiers is to establish a relation between what can be extracted from the videos, i.e., low-level multimodal features, and the event. Their performance is thus closely linked to their ability in finding the best combination of features and to derive a decision rule that allow discriminating the positive examples with an adequate level of generalization.

We focus here on probabilistic approaches to event detection which are, for the most part, based on the Bayesian theory. One of the interests of statistical models is that they allow to easily take into account the correlations between features and the temporal aspect of the videos. The variety of statistical models used in multimedia analysis includes naive Bayes classifiers, decision trees, Bayesian networks—often in a naive Bayes way—and hidden Markov models (HMMs). In particular, HMMs and their variants have been extensively used for event detection in videos, with a wide range of applications from commercial detection (Mizutani et al. 2005), video genre classification (Gibert et al. 2003) to structure analysis in videos (Kijak et al. 2006; Qian et al. 2011). Hidden Markov models are well adapted to dense segmentation where every shot correspond to some well-defined event but not particularly suited for sparse event detection. Moreover, the model assumes that all observations are part of a single observation vector. This assumption is not well adapted to videos, in which several modalities are combined. Multistream extensions of the HMM framework have been proposed in audiovisual speech recognition, with application to multimedia analysis (Huang et al. 1999; Kijak et al. 2006). However, recent work on segmental multistream HMMs for tennis video structuring (Delakis et al. 2008) demonstrated the necessity of modeling the dependencies between features. Unfortunately, knowing which dependencies to model, and how, is not an easy task and requires a lot of human expertise, if possible at all. This paper therefore focuses on the necessity for algorithms to learn statistical dependencies between

variables in a large set of variables, along with the corresponding model.

In this regard, Bayesian networks (BN) define a general framework for probabilistic modeling which encompasses all of the above mentioned models, including HMMs and segment models using so-called dynamic Bayesian networks (Murphy 2002). A Bayesian network, dynamic or not, is a directed acyclic graph (DAG) where nodes represent random variables and links represent the causal relations between variables, thus allowing for a wide variety of topologies and offering flexibility. Moreover, BNs have shown to be adapted to multimodal fusion in the framework of video analysis (Huang et al. 2006; Lakka et al. 2011) with application in event detection, for example in soccer videos (Wang et al. 2004) or in Formula 1 car races (Petkovic et al. 2002).

Certainly one of the most appealing features of the Bayesian network theory is the ability to learn the structure of the graph that links together all the variables considered. In other words, one can learn the structure of the DAG in addition to the parameters of the model, a fact never satisfactorily achieved with HMMs. The main interest for structure learning is to avoid resorting to human expertise and heavy trial and error experimental protocols to define the best statistical model, yet avoiding unnecessary assumptions on the data such as the state conditional independence assumption for HMMs.

Several algorithms for structure learning in BNs, such as the K2 algorithm (Cooper and Herskovits 1992), have been proposed in the literature. In particular, score based approaches seek to maximize an objective function that reflects a trade-off between the best fit of the training data and the generalization capabilities of the model. However, such algorithms are seldom used in the multimedia area where naive networks predominate (Nefian et al. 2002). Note however the work of Choudhury et al. (2002); Friedman et al. (2000); Baghdadi et al.

(2008) which investigate structure learning successfully on rather simple tasks.

In this paper, we investigate the use of structure learning algorithms for a rather complex multimedia task which consists in detecting action shots in soccer videos from multimodal input. To the best of our knowledge, this constitutes the first attempt in the multimedia area to use BN structure learning algorithms on a large scale complex task. We illustrate that classical score-oriented structure learning algorithms such as the K2 one, whose usefulness has been demonstrated on simple tasks, fail at providing a good network structure for classification tasks where many correlated observed variables are necessary to make a decision. We then compare several structure learning objective functions, which aim at finding out the structure that yields the best classification results, extending existing solutions in the literature. All structure learning algorithms are evaluated and compared on a realistic task, namely action detection in soccer videos from multimodal input, using a comprehensive data set of 7 games.

The paper is organized as follows. Section 2 defines Bayesian networks more formally and present classical structure learning strategies. In section 3, we exhibit a case where the K2 algorithm performs poorly and discuss the reasons for this. We then propose new objective functions oriented towards the classification goal in section 4 and provide an experimental comparative study in section 5.

2. Bayesian networks and structure learning

A Bayesian network can be seen as a graphical representation of a probabilistic distribution over a set of random variables, where the graphical representation depicts the causal

relations between the variables. As with any classification model, using Bayesian networks faces two issues: inference and training. Inference aims at making a decision (as to whether or not the event considered is present in our case) based on the evidences, or observations, available, given a network. Training includes two usually distinct phases, namely the design of the model and the estimation of the parameters from examples given the structure. We first formally define Bayesian network and briefly discuss the inference issue, before discussing model design issues and presenting the general principles of structure learning in Bayesian networks.

a. Definition and inference

Formally, a Bayesian network is a statistical model of a set of random variables, representing relations between variables such as conditional independence or causality. A network can be represented as a graphical model, i.e., a direct acyclic graph (DAG) $\mathcal{G}$ where each node X_i is a random variable, arcs representing a relation of conditional dependence between the two variables at stake. In other words, the arc $X_i \rightarrow X_j$ indicates that X_j depends on X_i. Assuming random variable taking values in a discrete observation space, a probability distribution table is associated to each node X_i of the DAG to describe the probability of the random variable taking a value conditionally on the value of the set $\mathcal{P}_i$ made of the parents of node X_i, i.e. $\mathcal{P}_i = \{X_j \text{ s.t. } X_j \rightarrow X_i\}$. Hence, the network encodes the relations within the set of random variables considered, $\{X_i\}$, and can be used to factor the total probability of the collection according to

$$P[X_0, X_1, \ldots, X_n] = \prod_{i=0}^{n} P[X_i|\mathcal{P}_i] \ , \tag{1}$$

where $\mathcal{P}_i$ denote the set of variables corresponding to the parents for node i, i.e., the set of random variables upon which X_i is dependent.

A simple example of a Bayesian network is illustrated in Fig. 1, where the three variables X_1, X_2, X_3 are all independent conditionally on the knowledge of X_0. The total probability can therefore be decomposed as

$$P[X_0, X_1, X_2, X_3] = P[X_0]P[X_1|X_0]P[X_2|X_0]P[X_3|X_0] \ .$$

As can be seen, a Bayesian network over a given collection of random variables is fully defined by its structure, i.e., the topology of the direct acyclic graph, and by the conditional probability tables (CPT) at each node. In practice, Bayesian networks are mostly used to make decisions based on the *inference* of the value of unobserved variables given the observed ones. For example, in the framework of multimedia content classification that we are studying, one is interested in inferring the class value given observations. This can be done using so-called *naive* structures as the one depicted in Fig. 1 where X_0 represents the unknown class to be inferred—here, a binary class stating whether an event is present or not—given observations X_1, X_2, X_3. Inference algorithms, such as Kim and Peal (1983) and Jensen et al. (1990), are the key to solving marginalization or posterior problems so as to find an optimal configuration for unobserved variables.

b. Graphical model design

Apart from the inference issue, model design is a crucial step in implementing a Bayesian network classifier. Model design can be seen as a two-step process where the first step consists in defining the topology of the model while the second one relates to the estimation

of the conditional probability tables from training data.

Maximum likelihood approaches have been designed for parameter estimation in a variety of networks, exploiting the factorization of the total probability in the network (Heckerman 1995; Ghahramani 1998; Murphy 2002). For simple networks such as the ones considered in this study, with discrete variables, all of them observable in the training data, maximum likelihood estimation boils down to estimating conditional probabilities with empirical frequencies.

On the contrary, only a few algorithms have been proposed for the estimation of an optimal graph topology given training data. Moreover, these algorithms are seldom used in practice for real-life classification problems such as the one targeted here. Approaches to structure learning can be grouped into two main approaches. The first one consists in using statistical dependency tests to search for causalities between variables, such as in the IC and SGS algorithms (Pearl and Verma 1992; Spirtes et al. 1993). The second family groups methods targeting the optimization of a score that evaluates the quality of a structure. Several scores have been proposed in the literature, along with efficient—but suboptimal—strategies to review a large set of candidate structures to choose from. For example, restricting the possible set of structures to trees, one can search for the best tree structure using the maximum weight spanning tree (MWST) algorithm (Kruskal 1956), assuming the availability of a causality score between any two variables (Chow and Liu 1968; Heckerman 1995). For structures more general than trees, most algorithms, including the popular K2 algorithm (Cooper and Herskovits 1992) and its variants, imposes an ordering on the nodes such that the set of possible parents for a given node is limited to those nodes with a higher rank, thus drastically reducing the search space. Finally, greedy search

heuristics were also proposed in the literature for efficient exploration of the space of all possible structures (Chickering et al. 1995).

Most score-oriented structure learning algorithms exploit a score which seeks for a trade-off between accurately modeling the training data and obtaining a low complexity network. In particular, a popular score function which derives from a simplification of the K2 algorithm, is the Bayesian information criterion (BIC) for which the objective function to optimize is given for a graph $\mathcal{G}$ over N variables by

$$\begin{aligned} Q_{\text{BIC}}(\mathcal{G}) &= \ln P_{\mathcal{G}}[\mathbf{X}] - \frac{\lambda}{2} C(\mathcal{G}) \ln(K) \\ &= \sum_{i=0}^{N} \left(\ln P[X_i|\mathcal{P}_i] - \frac{\lambda}{2} C_i(\mathcal{G}) \ln(K) \right) \end{aligned} \tag{2}$$

where K is the number of training examples, $\mathbf{X}$ is the set of variables, $C(\mathcal{G})$ is the total number of free parameters in the network and $C_i(\mathcal{G})$ is the number of free parameters in node i. As explicitly shown in the above equations, the objective function is decomposable as a sum over all nodes of the network, thus limiting the amount of computation to get a new score when the structure is changed.

However, this approach suffers from severe drawbacks in a complex classification task, in particular because of the fact that the objective function is oriented towards description of the data, rather than towards optimal prediction. Indeed, in classification networks, the classification node, denoted X_c in the sequel, plays a particular role and should be treated differently. Few criteria were proposed for the purpose of classification. The tree augmented network (TAN) consists in augmenting a naive network with a tree structure using a MWST algorithm (Geiger 1992; Friedman et al. 1997). More recently, Grossman and Domingo (2004) proposed to use the conditional likelihood—i.e., conditionally to the classification

node—rather the likelihood in the objective function.

In the next section, we study the use of the Bayesian information criterion as an objective function to learn the structure for the task of event detection in soccer videos and show the limitations of likelihood-based objective functions in this case. A classification-oriented objective function is proposed in section 4 and compared to the TAN algorithm and to a K2 augmented network.

3. Limitations of K2 structure learning for soccer video indexing

In preliminary work, we demonstrated the benefit of using Eq. 2 as the objective function for structure learning in the task of multimodal advertisement detection in videos (Baghdadi et al. 2008). Elaborating on these results, the same structure learning paradigm is here applied to a more complex task, where more variables are to be considered, namely the detection of actions in soccer videos based on low-level audio and visual features.

We first describe the task and experimental protocol that is used throughout the paper before presenting results which demonstrate that K2 fails at such a complex classification task.

a. The action detection task in soccer videos

We consider the task of detecting actions in soccer videos, where an action is defined as a period of time in the match when a player is about to shoot to score. Such an action

usually takes place near the goal mouth and comes with the cheering of the crowd and an excitement of the speaker, thus requiring multimodal input. Some replays of the action also usually follow. Action detection in soccer video is a complex task which requires that multiple features be considered simultaneously, thus being far more challenging for structure learning algorithms than previous case studies. In particular, in comparison with advertisement detection, the number of features required to accurately identify actions is greater and no straightforward features such as monochrome frames are available.

In this work, detection is performed at the shot level, all videos being automatically segmented into shots. From each shot, the following set of 8 binary audio and visual features is automatically extracted, a value of one indicating the presence of the feature:

i. crowd excitement: this feature usually is strongly related with a noticeable event;

ii. transition shot: some transition effects, classically detected by the shot segmentation algorithm, are usually added to increase the attractiveness of an event;

iii. wide shot: a shot is classified as wide based on the detection of green as the dominant color and on the detection of terraces in the background;

iv. lull scene: a lull scene, as opposed to a peak scene, corresponds to a game sequence where nothing special is happening and for which directors usually alternate between wide shots and other shots to maintain the dynamics of the video. The detection of such scenes is mostly rule based, relying on the result of the classification of shots into wide or not;

v. presence of face: shots containing mostly a face, as indicated by a face detection

algorithm, are likely to be close-up which are strongly related to action;

vi. green shot: shots where the green color is sufficiently present are marked as so to indicate whether the field is visible or not;

vii. replay logo: this feature indicates the presence of a replay logo which indicates the start or end of a replay sequence;

viii. goal mouth: this indicates whether the goal mouth is visible or not, thus acting as an indicator of the action importance.

However, actions are mostly characterized by the temporal evolution of the features and classification can hardly be performed on the base of a single shot. Hence, the primary features of a shot are augmented with features from the neighboring shots, taking 2 shots of context on the left and right side respectively. This amounts to a total of 40 contextual features which are to be modeled using a Bayesian network classifier. With respect to our previous study, it is important to note that this constitutes a larger set of features. Moreover, features are highly correlated, in particular due to the use of context shots. Finally, it should also be noted that not all features are always directly relevant for the classification task.

Experiments are carried out on a data set of 7 games broadcasted during the 2006 World Cup, amounting to about 14 hours of video. Automatic shot segmentation yielded 9,632 shots in total, among which 192 were labeled by a human expert as action shots (about 2 % of the total number of shots). Table 1 provides details on a per video basis. Due to the limited number of data available a cross-validation protocol with 7 folds was adopted, retaining one match as test material for each fold. The training set for each fold is used both for structure learning and maximum likelihood parameter estimation, the two being

performed jointly regardless of the structure learning criterion used. Results are reported in terms of recall and precision on the action shots, where a shot is deemed to be an action shot if the posterior probability $P(X_c/X_1, X_2, ...X_n)$ of the classification node is above a threshold. The threshold is varied so as to achieve different trade-offs between recall and precision.

b. K2 structure learning for soccer videos

Following exactly the same methodology as in (Baghdadi et al. 2008) where the K2 algorithm demonstrated effectiveness for advertisement detection, the K2 algorithm using the score function given in Eq. 2 was first used for event detection in soccer video. The structure was initialized using the MWST algorithm of Chow and Liu (1968), taking the classification node X_c as the root for the tree. Node ordering in the K2 algorithm was derived from the tree structure obtained from the initialization step.

Results are reported in Fig. 2 and compared with a naive structure. Contrarily to previous results on advertisement detection, the K2 structure fails at capturing the complex relations between variables, resulting in poorer performance than the naive structure. This result is counter-intuitive as one would expect the model to benefit from taking into account the correlations that might exist between variables. An example of a structure learned from the data is given in Fig. 3, where variables were assigned arbitrary numbers, and illustrates two important points regarding the behavior of the K2 algorithm. On the one hand, structure learning succeeds to some extent in capturing the relations between variables, resulting in a network structure rather different from the naive one. This structure nevertheless

remains difficult to interpret, even with some expert knowledge in soccer. But, most of all, it also appears that only a few feature (observed) nodes are directly connected to the event classification node (red links), contrary to the naive network where, by definition, all features are connected to X_c.

This last observation is the key to understanding the poor results obtained with structure learning when, contrarily to the advertisement detection use case, a large number of variables is at stake. Indeed, for a graph structure $\mathcal{G}$, the likelihood term in the score function Q_{BIC} can be rewritten as

$$\ln P_{\mathcal{G}}[\mathbf{X}] = \ln(P_{\mathcal{G}}[X_1, ..., X_n]) + \ln(P_{\mathcal{G}}[X_c|X_1, ..., X_n]) \ . \tag{3}$$

From this formulation, it can be seen that as the number n of observed variables increases, the first term on the right hand side of the equality decreases rapidly. Since the second term does not depend on n, structure learning with a large number of variables is dominated by the maximization of the term $\ln(P_{\mathcal{G}}[X_1, ..., X_n])$, regardless of the class node X_c. The result is a structure that represents the relationships between the observed variables, regardless of their impact on the classification task considered.

These preliminary results show that for classification tasks with a large number of observed variables, structure learning algorithms searching for a trade-off between the best fit of the data and the complexity of the resulting model are not suited. A solution to skirt this issue is feature selection so as to limit the number of observed variables, a solution that has proven experimentally valid. However, feature selection might result in information loss. An alternate solution consists in using objective functions for structure learning that account for the specificity of classification tasks, paying special attention to the peculiarity of X_c.

4. Classification-oriented structure learning algorithms

Two main strategies can be envisioned to learn the structure of a Bayesian network for classification. The first one consists in forcing relations between the observed variables and the classification node, leaving structure learning to the sole relations between observed variables. The second one consists in explicitly accounting for classification issues in the objective function. The first option benefits from an easy implementation but is suboptimal, while the second one is optimal but difficult to implement because most classification-oriented objective functions are not decomposable.

For each of the two strategies, an efficient algorithm for Bayesian network structure learning in the framework of classification is proposed. Firstly, pursuing the philosophy of the tree augmented network structure learning algorithm of Friedman et al. (1997), we impose constraints on the optimal structure, forcing all observed variables to be directly related to the classification node. This strategy yields a K2 augmented network structure which is still learned based on the likelihood-complexity trade-off. Secondly, a discriminative objective function (Grossman and Domingo 2004) is used in replacement of the K2 likelihood based one. Unfortunately, this new objective function is not decomposable over the set of nodes in the network and we resort to genetic algorithms for greedy optimization.

a. K2 augmented structure

As naive networks have proven successful for classification tasks on many an occasion, Friedman et al. (1997) proposed to augment the naive structure by adding arcs between observed variables using a MWST algorithm to generate a tree structure between the observed

nodes. A score based on the mutual information between each pair of nodes, conditionally on X_c, was used as input to MWST computation. This algorithm therefore results in a structure where each observed node has two parents: the classification node and another feature node.

Restricting the search of the structure to the set of trees clearly limits the complexity of the structure learning process. However, the relative simplicity of the resulting structure has also some drawbacks. First the tree structure will not allow to have connections between more than two features nodes, a fact that is likely to appear when a large set of variables is used. This is particularly true in our case because of the use of contextual features from the neighboring shots in the description of each shot, likely to be highly correlated one to another. Additionally, accounting for features not related to any other (i.e., a feature that should exhibit a unique connection to X_c) is impossible.

As a workaround, we propose to augment the naive structure using K2 structure learning with the BIC criterion, thus extending the tree augmented network philosophy. Using K2 structure learning to augment the naive structure clearly enlarges the set of possible structures, enabling more complex structures that do not suffer from the limitations stated. K2 augmented structure learning relies on a modified version of the Bayesian information criterion which accounts for the compulsory link between each feature and the event node. Formally, the objective function is defined as

$$Q^{(c)}_{\mathrm{BIC}}(\mathcal{G}) = \sum_{i=0}^{N} \left(\ln P(X_i|\mathcal{P}_i, X_c) - \frac{\lambda}{2} C(X_i, \mathcal{G}) \ \ln(K) \right) \ , \tag{4}$$

and remains decomposable, thus making it possible to use the same efficient exploration strategy based on node ordering as for the initial K2 algorithm described in section 3.

b. Discriminative objective function

Even if the classification goal is explicitly considered in Eq. 4, the rationale still consists in finding the structure that best fits the training data, subject to the simplicity of the structure. Given a large number of observed variables, this structure might still be dominated by the search for a good explanation of the relations between features rather than by the search for the structure that best classifies the data. We therefore propose a new structure learning criterion with the goal of directly maximizing the class conditional probability $P[X_c|X_1, ..., X_n]$ rather than the joint probability $P[X_c, X_1, ..., X_n]$. The use of the class conditional probability was introduced in Greiner et al. (2005) for parameter estimation and studied in Grossman and Domingo (2004) for structure learning using the BNC algorithm. We detail here a variant of the BNC algorithm using a genetic algorithm to explore the space of possible network structures.

As in Grossman and Domingo (2004), the objective function for structure selection is defined as

$$Q_{\mathrm{CLL}} = \sum_{i=1}^{N} \ln P_{\mathcal{G}}[X_c|X_1, ..., X_n] \ . \tag{5}$$

Unfortunately, the discriminative score is not decomposable and cannot be written as a sum of local scores calculated separately for each node. We therefore resort to a genetic algorithm in order to explore the set of possible structures. Genetic algorithms are iterative algorithms that require an initial structure as a starting point. From this initial structure, a set of candidate structures is generated by adding, reverting or deleting one single arc. The discriminative score is calculated for each of the structures resulting from these mutations and the one that maximizes the score—given maximum likelihood estimates of the parameters—

is then chosen as the starting point for the next iteration. The algorithm stops if none of the generated structures increases the score. The choice of the initial structure is crucial and will be discussed along with experimental results in the next section.

5. Experimental results

Recall *vs.* precision trade-off curves are plotted in Fig. 4 for the tree and K2 augmented networks as well as for the discriminant objective function. Performance for the naive network is also reported as a baseline. It should be noted that for the K2 augmented network, λ was experimentally set to 3 in Eq. 4.

Results show that the two augmented approaches clearly improve over the naive baseline network. Indeed, by forcing the classification node to be connected to all the feature nodes, these techniques build a structure which benefits from the whole feature information, as with the naive Bayesian network, while also taking into account correlations between the features themselves. Moreover, the K2 augmented method provides better results than the TAN approach for the classification task, due to the fact that the resulting network is less constrained in the K2 augmented case. This allows for more flexibility to take into account the correlations between several features when they exist, or on the contrary to avoid non-relevant connections between features when this is not needed, a fact that is crucial in such a complex classification task as the one we are targeting. Finally, the classification-oriented approach based on a discriminative objective function actually outperforms all techniques. The maximization of this new score, dedicated to the classification task, is one explanation for these good results, the other being the absence of format restriction in the choice of the

final structure. The resulting network will therefore better describe the correlations between the classification node and the features. An analysis of the resulting structure with this classification-oriented scheme is proposed in the following section.

An example of a structure obtained with the classification-oriented objective function is shown in Fig. 5, where node 1 corresponds to the classification node. An analysis of this structure highlights the few number of nodes used for the classification process: 13 nodes are directly connected to the classification node (red links) and 10 additional nodes are indirectly used. A total of 17 nodes was therefore rejected from the final structure, as non relevant for the classification task at hand. Connections with these nodes were indeed not increasing the discriminative score and the algorithm therefore applies implicit feature selection. It is interesting to note that implicit feature selection cannot, by construction, occur in augmented techniques. This reduction of the structure size consequently results in a more reliable parameter learning step, hence in an increased performance for the classification.

As mentioned previously, the choice of the initial point for the genetic exploration of the set of possible structures for the conditional probability maximization is usually of utmost importance. Three initial points were tested, namely naive, TAN and K2 augmented networks. We observed no impact on performance after structure learning. However, training time is significantly affected by the initialization point, as reported in Tab. 2. The much reduced training time when starting from one of the augmented structure highlights the quality of the latter, thus making good starting point for the genetic algorithm.

6. Conclusion

Taking action detection in soccer videos as a use case for multimodal event detection, we have shown how structure learning in Bayesian networks, associated with the adequate objective function, can efficiently detect complex multimodal events in videos. We have demonstrated that, while using an information criterion for structure learning in BNs suffers major drawbacks for complex classification tasks with a large number of correlated observed variables, classification-oriented objective functions can efficiently deal with such matter. In particular, we proposed a new K2 augmented network structure and a genetic implementation of the conditional likelihood objective function which both turned to outperform state-of-the-art structure learning methods. Experimental results however call for a few remarks and suggestions for further work.

Firstly, we observed that the ability to select relevant variables, i.e., to decide that some variable has no direct or indirect relation with the classification node, is crucial. While the conditional likelihood criterion embeds feature selection, it is not the case for the BIC and augmented approaches which might benefit from explicit feature selection. However, we believe that embedding structure learning and feature selection is better suited than performing feature selection as a required preliminary step to structure learning.

Secondly, we considered that the observed variables directly contribute to the classification if at all. In other terms, we only have two types of variables, the observed ones and the classification node. For complex classification tasks, it appears interesting to consider hidden variables which act as intermediate concepts between the observation and the decision. We are convinced that training such networks will help improve the structure inferred and

thus classification by summarizing complex information from the features in a few concepts. However, extensions of the existing structure learning algorithms to handle hidden variables are required to do so.

Finally, the temporal dimension of videos was limited in this work to the use of contextual features from the neighboring shots, classification being performed on a per shot basis. As for shot-based classification, learning the temporal structure of the video with a goal-oriented objective function is likely to improve classification performance and requires further investigation. Directly learning the structure of a dynamic BN is intractable, except in some rare cases. Combining BNs and segmental HMMs appears like a plausible alternative, where the structure and parameter of BNs are trained to predict the posterior probabilities required for Viterbi decoding in segmental HMMs.

Acknowledgments

This work was partially funded by OSEO, French state agency for innovation, in the framework of the Quaero project.

REFERENCES

Bae, T. M., C. S. Kim, S. H. Jin, K. H. Kim, and Y. M. Ro, 2005: Semantic event detection in structured video using hybrid HMM/SVM. *Intl. Conf. on Image and Video Retrieval*,

113–122.

Baghdadi, S., G. Gravier, C.-H. Demarty, and P. Gros, 2008: Structure learning in Bayesian network based video indexing. *IEEE Intl. Conf. on Multimedia and Exhibition*, 667–680.

Chickering, D., D. Geiger, and D. Heckerman, 1995: Learning Bayesian networks: Search methods and experimental results. *Conf. on Artificial Intelligence and Statistics*, 112–128.

Choudhury, T., J. M. Rehg, V. Pavlovic, and A. Pentland, 2002: Boosting and structure learning in dynamic Bayesian networks for audio-visual speaker detection. *IEEE Intl. Conf. on Pattern Recognition*.

Chow, C. and C. Liu, 1968: Approximating discrete probability distributions with dependence trees. *IEEE Transaction on Information Theory*, **11 (3)**, 462–467.

Cooper, G. F. and E. Herskovits, 1992: A Bayesian method for the induction of probabilistic networks from data. *Machine learning*, **9 (2)**, 309–347.

Delakis, M., G. Gravier, and P. Gros, 2008: Audiovisual integration with segment models for tennis video parsing. *Computer Vision and Image Understanding*, **111 (2)**, 142–154.

Friedman, N., D. Geiger, and M. Goldszmid, 1997: Bayesian network classifiers. *Machine Learning*, **29 (2)**, 131–163.

Friedman, N., M. Linial, I. Nachman, and D. Peer, 2000: Using Bayesian network to analyze expression data. *Journal of Computational Biology*, **7 (3-4)**, 601–620.

Geiger, D., 1992: An entropy-based learning algorithm of Bayesian conditional trees. *Conf. on Uncertainty in Artificial Intelligence*, 92–97.

Ghahramani, Z., 1998: Learning dynamic Bayesian networks. *Adaptive Processing of Sequences and Data Structures*, C. Giles and M. Gori, Eds., Springer-Verlag, Lecture Notes in Artificial Intelligence, 168–197.

Gibert, X., H. Li, and D. Doermann, 2003: Sports video classification using HMMs. *IEEE Intl. Conf. on Multimedia and Exhibition*, 345–348.

Greiner, R., X. Su, B. Shen, and W. Zhou, 2005: Structural extension to logistic regression: Discriminative parameter learning of belief net classifiers. *Machine Learning Journal*, **59 (3)**, 297–322.

Grossman, D. and P. Domingo, 2004: Learning Bayesian network classifiers by maximizing conditional likelihood. *Intl. Conf. on Machine Learning*, 46–53.

Haering, N., R. Qian, and M. Sezan, 2000: A semantic event-detection approach and its application to detecting hunts in wildlife video. *IEEE Transactions on Circuits and Systems for Video Technology*, **10 (6)**, 857–868.

Heckerman, D., 1995: A tutorial on learning with Bayesian networks. Tech. Rep. MSR-TR-95-06, Microsoft Research.

Huang, C.-L., H.-C. Shih, and C.-Y. Chao, 2006: Semantic analysis of soccer video using dynamic bayesian network. *IEEE Trans. on Multimedia*, **15 (10)**, 1225–1233.

Huang, J., Z. Liu, Y. Wang, Y. Chen, and E. K. Wong, 1999: Integration of multimodal features for video scene classification based on HMM. *Wokshop on Multimedia Signal Processing*, 53–58.

Jensen, F. V., S. L. Lauritzen, and K. G. Olsen, 1990: Bayesian updating in recursive graphical models by local computations. *Computational Statistics Quarterly*, **4**, 269–282.

Kijak, E., G. Gravier, L. Oisel, and P. Gros, 2006: Audiovisual integration for tennis broadcast structuring. *Multimedia Tools and Application*, **30 (3)**, 289–311.

Kim, J. H. and J. Peal, 1983: A computational model for causal and diagnostic reasoning in inference systems. *Intl. Joint Conf. on Artificial Intelligence*, 190–193.

Kruskal, J. B., 1956: On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. of the American Mathematical Society*, Vol. 7, 48–50.

Lakka, C., S. Nikolopoulos, C. Varytimidis, and I. Kompatsiaris, 2011: A Bayesian network modeling approach for cross media analysis. *Signal Processing: Image Communication*, **26 (3)**, 175–193.

Lienhart, R., S. Pfeiffer, and W. Effelsberg, 1998: Scene determination based on video and audio features. *Multimedia, Tools and Applications*, **15 (1)**, 59–81.

Mizutani, M., S. Ebadollahi, and S. Chang, 2005: Commercial detection in heterogeneous video streams using fused multi-modal and temporal features. *IEEE Intl. Conf. on Acoustic, Speech and Signal Processing*, 157–160.

Murphy, K., 2002: Dynamic Bayesian networks: representation, inference and learning. Ph.D. thesis, University of California, Berkeley.

Nefian, A. V., L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, 2002: A coupled

HMM for audio-visual speech recognition. *IEEE Intl. Conf. on Acoustic, Speech and Signal Processing*.

Pearl, J. and T. Verma, 1992: A statistical semantics for causation. *Statistics and Computing*, **2 (2)**, 91–95.

Perlovsky, L. I., 1998: Conundrum of combinatorial complexity. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **20 (6)**, 666–670.

Petkovic, M., V. Mihajlovic, W. Jonker, and S. Djordjevic-Kajan, 2002: Multi-modal extraction of highlights from TV Formula 1 programs. *IEEE Intl. Conf. on Multimedia and Exhibition*, 817–820.

Qian, X., G. Liu, H. Wang, Z. Li, and Z. Wang, 2011: Soccer video event detection by fusing middle level visual semantics of an event clip. *Advances in Multimedia Information Processing - PCM 2010*, G. Qiu, K. Lam, H. Kiya, X.-Y. Xue, C.-C. Kuo, and M. Lew, Eds., Springer Berlin / Heidelberg, Lecture Notes in Computer Science, Vol. 6298, 439–451.

Saraceno, C. and R. Leonardi, 1998: Identification of story units in audio-visual sequences by joint audio and video processing. *IEEE Intl. Conf. on Image Processing*, 363–367.

Snoek, C. and M. Worring, 2005: Multimedia event-based video indexing using time intervals. *IEEE Transactions on Multimedia*, **7 (4)**, 638–647.

Spirtes, P., C. Glymour, and R. Scheines, 1993: *Causation, Predection and Search*. Springer-Verlag.

Tovinkere, V. and R. Qian, 2001: Detecting semantic events in soccer games: Towards a complete solution. *IEEE Intl. Conf. on Multimedia and Exhibition*, 833–836.

Wang, F., Y.-F. Ma, H.-J. Zhang, and J.-T. Li, 2004: Dynamic Bayesian network based event detection for soccer highlight extraction. *IEEE Intl. Conf. on Image Processing*, 633–636.

Xu, G., Y.-F. Ma, H.-J. Zhang, and S. Yang, 2002: Motion based event recognition using HMM. *IEEE Intl. Conf. on Pattern Recognition*, 831–834.

Zhong, D. and S. Chang, 2001: Structure analysis of sports video using domain models. *IEEE Intl. Conf. on Multimedia and Exhibition*, 713–716.

List of Tables

Game	#shots	#action shots	duration of video
Germany – Argentina	1,983	20	2h52
Saudi Arabia – Ukraine	1,234	27	2h01
France – Brasil	1,736	30	2h10
Germany – Italy	1,490	38	2h33
Italy – Ukraine	1,181	23	1h52
Germany – Portugal	694	17	1h00
Brasil – Ghana	1,122	37	1h56
Total	9,440	192	14h24

TABLE 1. Number of shots, number of action shots and total duration of the video per game. Note that the Germany – Portugal video was truncated but kept as is.

Initial structure	Learning time
naive	2d : 4h : 50min
TAN	1d : 6h : 23min
K2 augmented	1d : 4h : 37min

TABLE 2. Impact of the initial structure on the learning time of the discriminative approach.

List of Figures

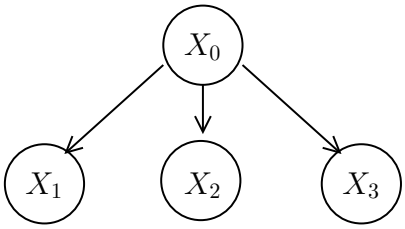

FIG. 1. Example of a simple Bayesian network with four variables.

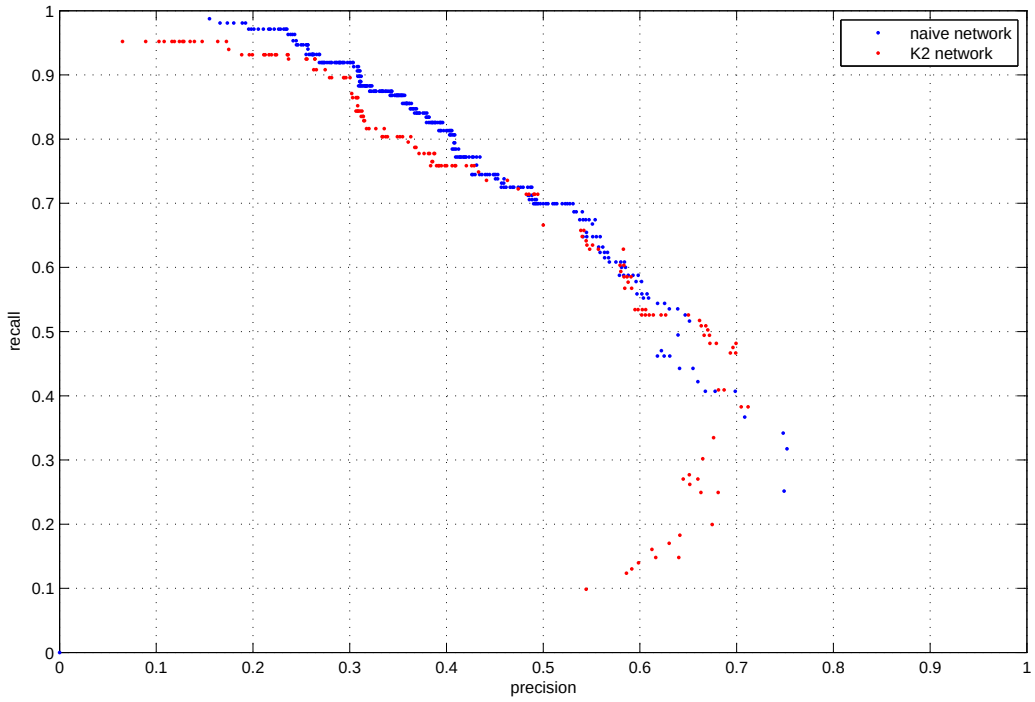

FIG. 2. Recall *vs.* precision trade-off curves comparing the K2 (red) and naive (blue) structures.

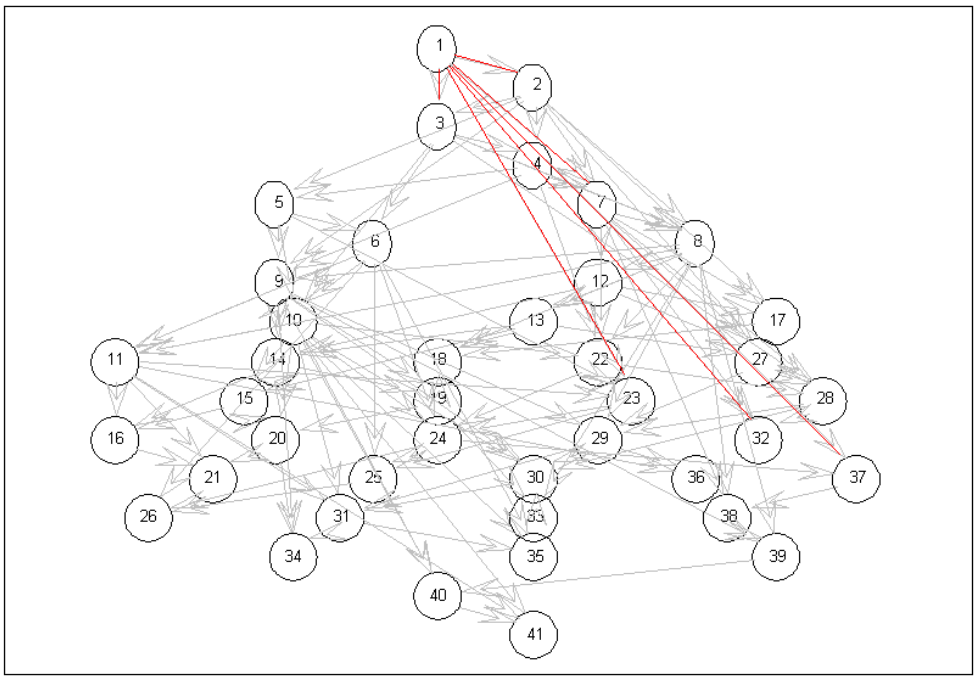

FIG. 3. Structure obtained with K2 structure learning, where red links denote direct connections between X_c (node 1 in the picture) and the observed variables.

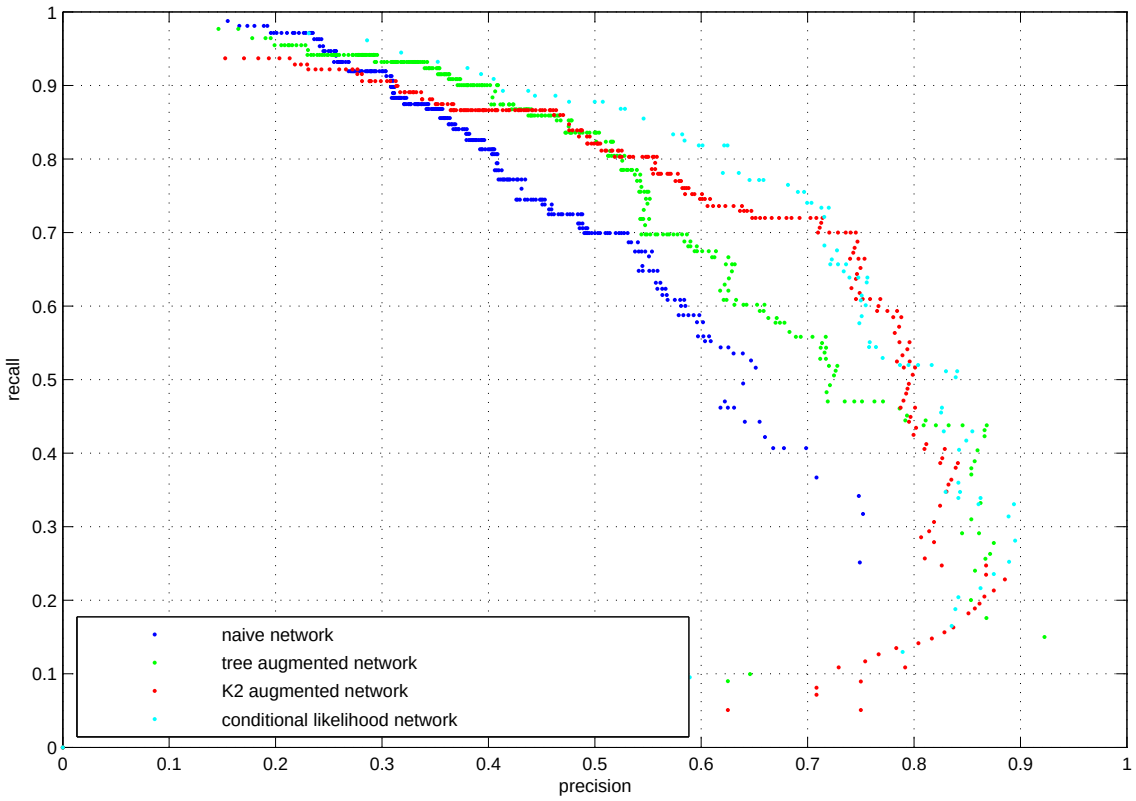

FIG. 4. Recall precision trade-off curves for i) a naive Bayesian network (dark blue), ii) a tree augmented network (green), iii) a network resulting from the K2 augmented technique (red) and, iv) a discriminatively trained network (light blue).

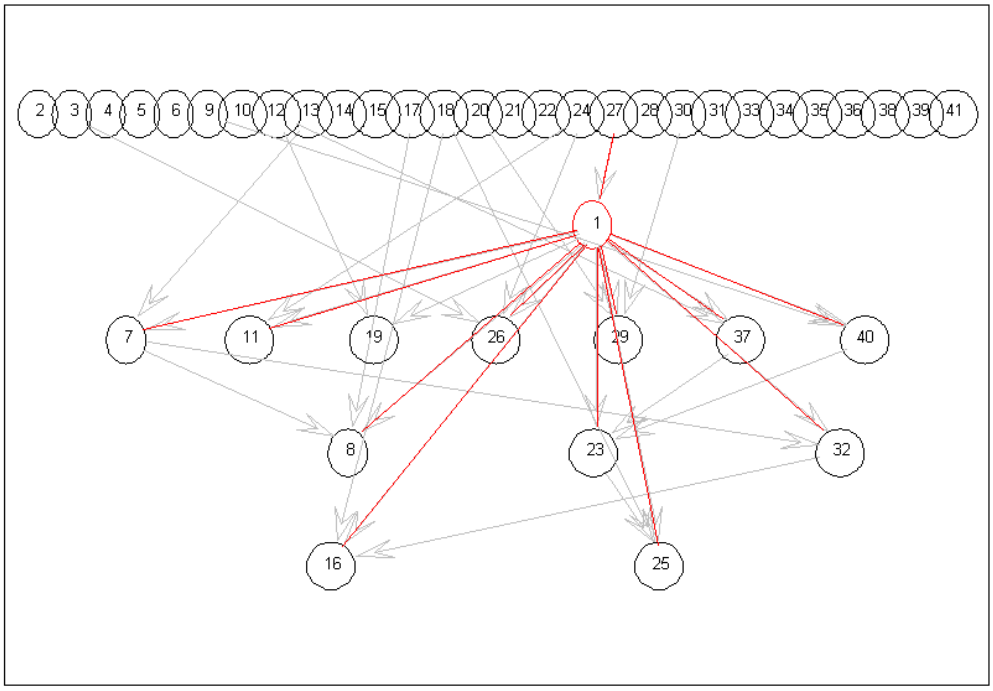

FIG. 5. Example of a structure resulting from the use of the discriminative objective function.